Lucky Jordan

Project Overview

I downloaded text from Project Gutenberg in order to analyze its sentiment and word frequency. I used pickling to get the the text and then manually copied and pasted a particular excerpt to a text file. I then used nltk (word_tokenize and stopwords), vaderSentiment, Counter, numpy, and matplotlib to preprocess the text, analyze its sentiment, find the most common words, and plot them in a histogram.

I hoped to learn about preprocessing of text and finding word frequency of significant words in a large body of text. I accomplished both of these goals in this project.

Implementation

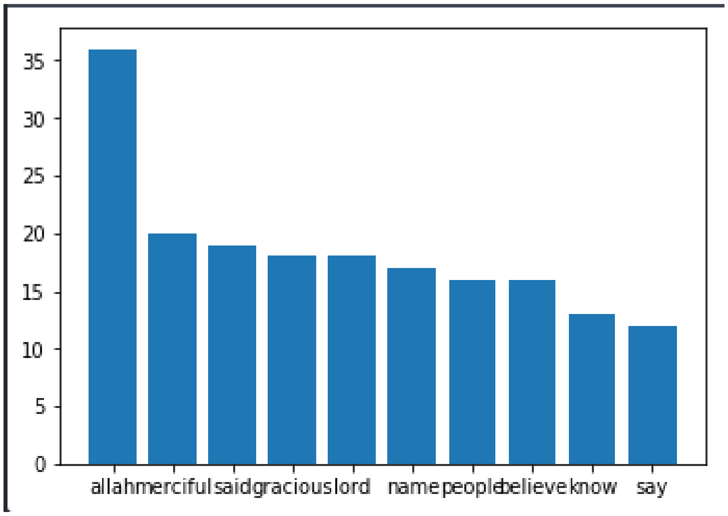
The major steps in my program were sentiment analysis, preprocessing, tokenization, frequency analysis, and creating a histogram. The sentiment analysis was straightforward and simple using vaderSentiment. The preprocessing took the text as a string and removed all characters that were not in a string which contained all English letters and the space character. This filtered text string was then passed to a function which used nltk to tokenize the string, emove stopwords (common and uninteresting words which have no value in being included in the histogram) and return a list of all the important or interesting words in the text.

I then used a funciton to plot the histogram of the most common words using Counter to create a list of tuples which contained the words and their frequencies. It then plotted this data using numpy and matplotlib. I could have instead used an ordered dictionary and a histogram function to generate an ordered list of the most common words. After finding Counter, it appeared that I could accomplish all that in one line of code as opposed to multiple functions with several loops.

Results

I found that the language was about equally negative and positive with the overwhelming majority being neutral. This data would be more interesting if compared to other religious texts or compared across sections of the Koran. Results = compound: -0.9875, neu: 0.786, neg: 0.111, and pos: 0.103.

The histogram represents the most common words. Some of the particular interesting ones which seem to possibly contradict the sentiment analysis are merciful and gracious. I believe this speaks to the quality of religious teachings in the Koran and how they relate to interpersonal interactions and care for humanity.



Reflection

I wish that I had kept my code more organized throughout the process and determined the major functions before starting. In addition, I wish I had committed my changes throughout the project so that I could refer to old versions without just copying files and deleting them later or commenting out a bunch of lines. I also wish that I had known about Counter before starting the project and spending time on generating a histogram from a dictionary.